Toward RNN Based Micro Non-verbal Behavior Generation for Virtual Listener Agents

Hung-Hsuan Huang[1,2(✉)], Masato Fukuda[1], and Toyoaki Nishida[1,2]

[1] RIKEN Center for Advanced Intelligence, Kyoto, Japan
[2] Graduate School of Informatics, Kyoto University, Kyoto, Japan
hhhuang@acm.org

Abstract. This work aims to develop a model to generate fine grained and reactive non-verbal idling behaviors of a virtual listener agent when a human user is talking to it. The target *micro* behaviors are facial expressions, head movements, and postures. The following two research questions then emerge. Whether these behaviors can be trained from the corresponding ones from the user's behaviors? If the answer is true, what kind of learning model can get high precision? We explored the use of two recurrent neural network (RNN) models (Gated Recurrent Unit, GRU and Long Short-term Memory, LSTM) to learn these behaviors from a human-human data corpus of active listening conversation. The data corpus contains 16 elderly-speaker/young-listener sessions and was collected by ourselves. The results show that this task can be achieved to some degree even with the baseline multi-layer perceptron models. Also, GRU showed best performance among the three compared structures.

Keywords: Virtual agent · Facial expression · Facial Action Coding System (FACS) · Multimodal interaction · Deep learning · Recurrent neural network (RNN) · Long Short-term Memory (LSTM) · Gated Recurrent Unit (GRU)

1 Introduction

The population of elderly people is growing rapidly in developed countries. If they do not maintain social life with others, they may feel loneliness and anxiety. For their mental health, it is reported effective to keep their social relationship with others, for example, the conversation with their caregivers or other elderly people. There are already some non-profit organizations recruiting volunteers for engaging "active listening" with the elderly. Active listening is a communication technique that the listener listens to the speaker carefully and attentively. The listener also ask questions for confirming or showing his/her concern about what the speaker said. This kind of support helps to make the elderly feel cared and to relieve their anxiety and loneliness. However, due to the lack of the number of volunteers comparing to that of the elderly who are living alone,

G. Meiselwitz (Ed.): HCII 2019, LNCS 11578, pp. 53–63, 2019.
https://doi.org/10.1007/978-3-030-21902-4_5

the volunteers may not be always available when they are needed. In order to improve the results, always-available and trustable conversational partners in sufficient number are demanded.

The ultimate goal of this study is the development of a computer graphics animated virtual listener who can engage active listening to serve elderly users at a level close to human listeners. In order to conduct successful active listening, it is considered essential for the listener to establish the rapport from the speaker (elderly users). Rapport is a mood which a person feels the connection and harmony with another person when (s)he is engaged in a pleasant relationship with him/her, and it helps to keep long-term relationships [8,9]. In order to realize this, like a human listener, the virtual listener has to observe the speaker's behaviors, to estimate how well the speaker is engaging the conversation [14], and then reacts to the user. The utterances of the speaker are obvious cues for the estimation of the speaker's engagement. However, due to the nature of active listening conversation, the speaker may utter in arbitrary contexts, It is difficult to utilize this information. Non-verbal behaviors are considered more general and more robust (less user-dependent). Previous works in generating the non-verbal behaviors of virtual agents usually adopt manually defined or machine learning rules to trigger predefined animation sequences [9]. When there is no attentional behavior being triggered, the character will stay steady or play so-called idling motions in a loop. Due to the fact that human body can never keep steady and always move slightly, though the movement may be neither meaningful nor attentional. Idling movements which are randomly generated or a looped replay of motion captured human movements are adopted. However, the character animation is still a repetitive replaying of exactly identical sequences or is not reactive to the user's behavior. These *repeated patterns* make the agent be perceived unnatural.

This work aims to develop a model to generate fine grained and reactive non-verbal behaviors of the virtual character when the human user is talking to it. The target *micro* non-verbal behaviors are facial expression, head movements, and postures. Then the following research questions emerge:

- Are the listener's behaviors learnable only from the speaker's behaviors? That is, the listener's behavior is only (mostly) dependent on the speaker.
- If the answer for the question above is true, which machine learning model will be appropriate? That is, which kind of regression model can achieve high precision in the learning task.

Deep neural networks have been proven to be effective in various learning tasks. Facial expressions, head movements, and postures involve dozens of parameters simultaneously, this is supposed to be an appropriate application for neural networks, which can generate multiple outputs in nature. In this paper, we presents our exploration on the use of recurrent neural networks (RNN) which is designed to capture time series data to learn the reactive behaviors to the human communication interlocutor's corresponding micro non-verbal behaviors. The generated animation is expected to be fine-grained both temporally and

spatially, no identical sequences, and reactive to the user's behaviors. We compared two typical temporal models, Long Short-term Memory (LSTM) [6] and a simpler model, Gated Recurrent Unit (GRU) [3] which omits the forget and output gates with the classic non-temporal multi-layer perceptron as a baseline. This paper is organized as the follows: Sect. 2 introduces related works, Sect. 3 introduces the data corpus used in the machine learning experiment, Sect. 4 describes the neural network models compared, and Sect. 5 concludes the paper.

2 Related Works

Deep neural networks (DNNs) have been shown their performance in generalizing the learning process of complex contexts including the multimodal classification of human behaviors [1]. Despite of the time consuming process in the learning phase of DNNs which usually requires large dataset, the application of learned models is fast. Therefore, DNN is a suitable tool for the generation of multiple parameters with large number of input variables. DNNs have been shown their power in image or speech recognition and are also gradually getting popular in human-agent interaction and social computing fields. They are most often used in prediction and estimation of human state such as visual focus [12] or sentiment [2] by using low-level multimodal features. Recently, DNNs are also seen in the generation part such as utterance [15] and gesture [5]. A Generative Adversarial Networks (GAN) is used in generating photo-realistic facial expression images in reacting to another facial expression [10]. However, there is no direct previous work in generating facial expression parameters to animate virtual characters in active listening context yet.

3 Active Listening Data Corpus

3.1 Data Collection Experiment

In order to collect data closer to the situation as talking with a virtual agent on screen, the data corpus was collected in with tele-communication sessions via Skype. For elderly people (69 to 73 years old, 71 in average) were recruited for the speaker roles, and four college students (averagely 22 years old) were recruited for listener roles. The genders of the participants were balanced in each age group. Only two elderly subjects knew one young subject while the other subjects met each other for the first time. Each elderly participant talks with every young participants for at least 15 min. The experiment recorded 16 sessions of dyadic conversation with length up to 30 min. All participants are native Japanese speakers and the conversation was done in Japanese.

There was no determined topic in the conversation, and the participants could talk freely. The subjects were instructed to talk as they met for first time. The listener participants were instructed to actively listen to the speakers, that is, instead of talking about themselves, they should talk about the elderly speakers, try to motivate the disclosure of the speakers, and let them enjoy the talks. The

results were, the two interlocutors in the conversation did not play equal roles, and the conversation sessions were usually conducted in interview style, i.e. the listener asks questions and the speakers answer them.

The monitor used at the elderly participant side is a large size TV (larger than 50″) while the video at young participant side is projected to a 100″ screen. In addition to the WebCams for Skype connection, two digital video camera with full HD resolution (1920 × 1080) at 29.97 fps were used to capture the participants from their front sides, camera positions were adjusted to cover their whole upper bodies (Fig. 1).

Fig. 1. Video recording of the active listening experiment. For the situation closer to face-to-face conversation, large size screens are used and the height of sitting position and screen is adjusted to align the heights of the subjects' eyes

3.2 Data Preparation

The video taken by the two video cameras were used for the extraction of multimodal features. Since the objective is the generation of listening behaviors while the speaker is talking, it is necessary to identify the time periods when the speaker is speaking. The behaviors of the speaker in those periods are then extracted as explanatory variables, and the behaviors of the listener in corresponding periods are extracted as response variables. Speakers' speech activities were automatically identified by the annotation software, ELAN [11] with additional manual corrections. The speaker of each utterance is manually labeled. Considering the balance of data length of each session, at most 20 min from the beginning are used in longer sessions.

The participants' facial expressions are extracted with an open source tool, OpenFace[1]. OpenFace estimates head posture, gaze direction, and 17 of 46 facial action units (AU) in accordance with Ekman's Facial Action Coding System (FACS) [4]. The posture information were extracted by using the open source tool, OpenPose[2]. Since OpenPose only generates two-dimensional coordinates

[1] https://github.com/TadasBaltrusaitis/OpenFace.
[2] https://github.com/CMU-Perceptual-Computing-Lab/openpose.

of the joints of human bodies, posture information (leaning in two axes, forward/backward and left/right) are approximated with the assumption that the widths of the participants' shoulder are the minimum values when they are sitting straight up (i.e. they only lean forward but backward). Prosodic features of speakers' voice were extracted by using the open source tool, OpenSMILE[3]. The 16 low-level descriptors (LLD) of the Interspeech 2009 Emotion feature set [13] were extracted at 100 fps. The features include root-mean-square of signal frame energy, zero-crossing rate of time signal, voicing probability, F0, and MFCC. All of the feature values are normalized to be within the range between 0.0 and 1.0.

Table 1 shows the overview of the prepared dataset. Despite the frames with partially invalid or missing data, totally there is four hours and 21 min of recorded data. Male listeners spoke less than female listeners. This coincides to the observation during the experiment, the two male listener participants performed relatively worse than the two females one. They were less skillful in motivating the speakers to talk more. Usually asked typical questions one by one and did not widen the topics in the answers from the speaker.

Table 1. Overview of the dataset. Data size is the set with two-second window

Listener	Speaking	Frame	Length (s)	Data size
Male	Yes	69,818	2,334	1.4 GB
	No	169,387	5,663	13.6 GB
Female	Yes	82,261	2,750	1.7 GB
	No	147,165	4,920	11.8 GB

4 RNN Models for Behavior Generation

4.1 Experiment Procedure

We formalize the purpose of this work as a regression problem from the speaker's facial expression (17 variables), gaze (8 variables), head movements (6 variables), and prosodic information of voice (16 variables) to the listener's (i.e. the agent) facial expression, head movements, and postures (2 variables). Since human behaviors are continuous activities of these parameters, the data frames are not independent to each other.

The dynamics of human behaviors are supposed to be better captured as time series data. Recurrent neural networks (RNN) which take the influence of previous input data into account are proposed for processing such time series data. Two variations of RNNs, Long Short-term Memory (LSTM) and Gated Recurrent Units (GRU), and the baseline multi-layer perceptron (MLP) were explored in this work. We designed a simple common three-layer network structure to run the evaluation experiments (Fig. 2). The first layer is separated into

[3] https://www.audeering.com/what-we-do/opensmile/.

two groups, one handles the input of video information (OpenFace), and the other one handles the input of audio information (OpenSMILE). The inputs are fed into LSTM/GRU layer with eight times units than the input, they then go trough a fully connected (dense) layer with 128 nodes, respectively. This absorbs the different frame rates in different modalities (30 fps v.s. 100 fps) and separates the temporal process of each modalities. The first fully connected layer has the same amount of nodes for each input group, this balances the influential power of the two input modalities (video and audio). They are then concatenated and go trough an additional fully connected layer that has 512 nodes before the output layer.

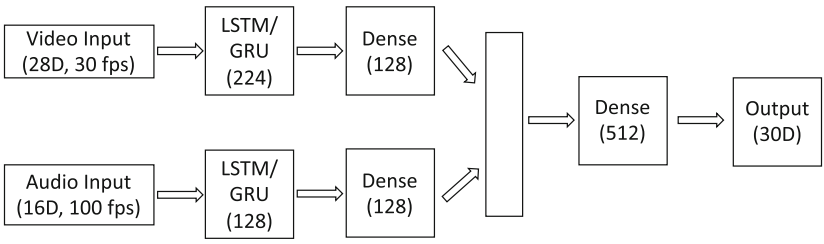

Fig. 2. Network architecture of the evaluation experiments

The same network structure is evaluated with LSTM and GRU units running on one-second and two-second long input data sequences. The raw data are transformed to trunks of time series with sliding window. In the cases of MLP models, the data trunks are reshaped to one-dimension arrays and are fed to the network in trunks by trunks. Therefore, the inputs for MLP with two seconds of data window will have $28 \times 30 \times 2 + 16 \times 100 \times 2 = 4{,}880$ dimensions. In addition to one-second and two-second windows, MLP was tested with zero-second window (i.e. the data of one frame only), too. The models are in seven combinations of the dataset, the whole dataset, male listeners, female listeners, and four individual listeners. All models are evaluated by leave-one-speaker cross validation, that is, the data when the listener(s) is/are interacting with one of the four speakers are extracted as the test dataset while the data of the other three speakers are used as training set. This procedure is repeated for four times so that the data of each speaker is used once as the test dataset. Then the average of the four trials is used as the final results. The experiment program is developed to use GPU computation where large amount of computation is done in parallel so that the reproducibility cannot be secured. Therefore, each experiment trial above was run for three times and the results are averaged. Each trial is trained for 200 epochs and the intermediate model which has best performance (lowest mean-squared-error (*mse*) upon the training dataset) is used for cross validation. *mse* is also adopted as the evaluation metric of the experiment. Since all data values are normalized to the range between 0 and 1, *mse* values can be interpreted regarding to this range.

4.2 Experiment Results

Experiment results are shown in Fig. 3. According to the evaluation results, we have the following findings:

- *mse* values of all temporal models except one-second LSTM one for male listeners are under 0.02. Considering the value range is up to 1.0, the errors are relatively low. This implies that the micron non-verbal behaviors of the listeners are indeed reactive to those of the speakers and therefore can be learned.
- The responsive behaviors are person dependent, individual models always perform better than gender-specific models and the whole-set model in the same setting. Among these, the models learned from female listeners generally perform better than male ones. This is probably because the female listeners were more skillful in communication, their reactions were more dynamic, and hence convey more characteristics which could be learned. On the contrary, male listener's reactions were more monotonous and were more difficult to be learned.
- GRU models generally perform better than corresponding LSTM ones on our dataset. Since GRU units have less parameters to be tuned, this may indicate that LSTM's additional parameters are over-killing and increase the difficulty of the learning task.
- As expected, MLP models generally perform worst, but surprisingly they still can be trained to some degree of precision even with the one-frame datasets.
- All models converges within 150 epochs, and there are no obvious impacts on the number of epochs in cross validation even though more epochs (we tried up to 500 epochs) tend to generate more precise models on the training set.
- Not always but two-second window size models often perform worse than their one-second counterparts. This may imply the listeners' reaction to the speaker does not require long period of perception (within one second). However, when actually apply the models to a realtime working agent system [7], the long-window version generates smoother (with less jittering) character animation from observation.
- Although there are only slight differences, two-second models perform better than one-second models on male listeners' datasets. On the other hand, one-second models perform considerably better than two-second ones on female listeners' datasets. This implies the different tendency in reacting to communication interlocutors for male and female subjects in our dataset.

Due to inherently different network structure, direct and completely fair comparison between the feature sets of different modalities is impossible in the case of neural networks. Out network settings were an approximation of comparable networks of multiple combinations of feature modalities and may not be the perfectly fair settings. Figure 4 depicts the comparison on the effectiveness of different combinations of modality features. Contrary to the priori expectation, multimodal models do not always perform better than uni-modal models. For

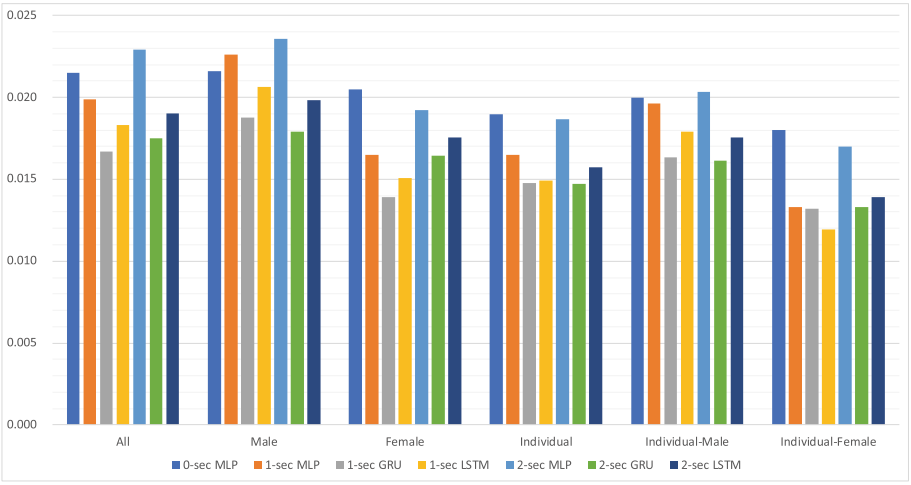

Fig. 3. Mean-squared-error values of compared models. "Individual" denotes the average of all individual listeners while "Individual-Male" and "Individual-Female" denotes the average values of male and female listeners, respectively

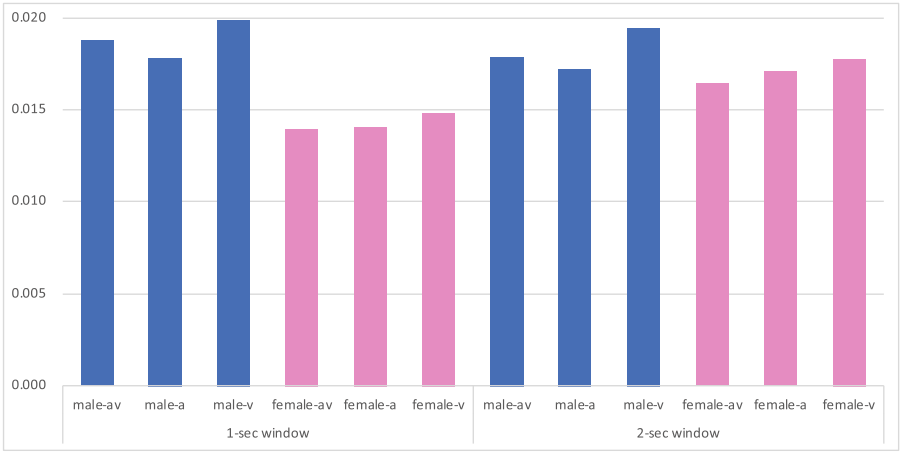

Fig. 4. Comparison of the leave-one-speaker-out cross validation results among different combinations of input modalities to GRU models using male and female listener datasets

male datasets, audio features perform best while the multimodal models perform best for female datasets. Comparing to video feature set, audio feature set is always more effective in cross validation measurements. On the other hand, from the learning curves depicted in Fig. 5, contrary to cross-validation results, video information is more active in network training itself. Video-only models converge faster and to a stable level with less errors. Jointly consider cross-

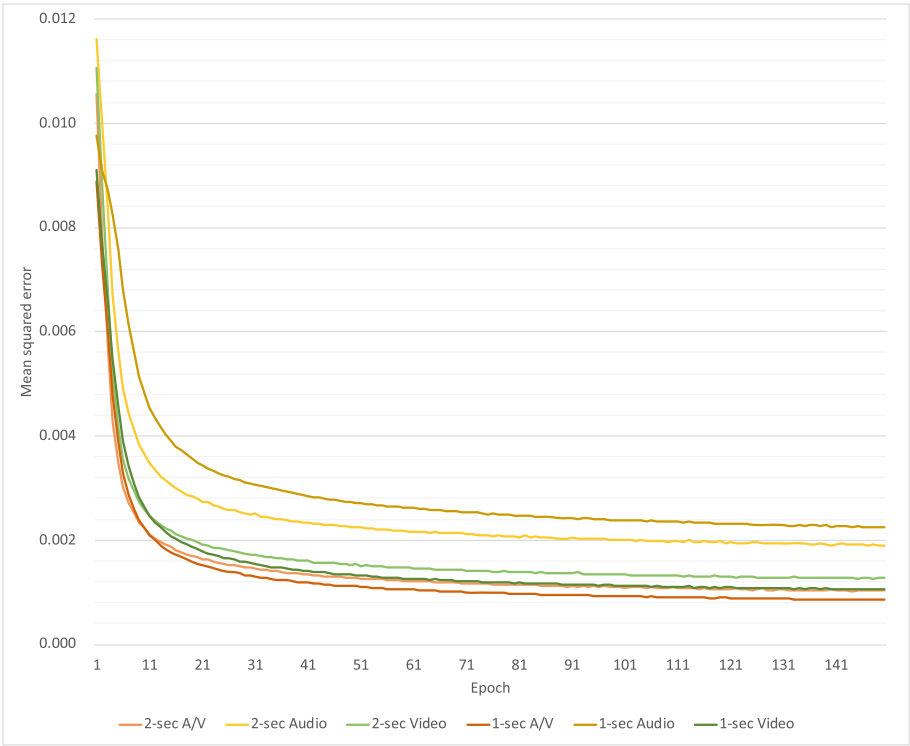

Fig. 5. Learning curves of the GRU models conducted from female listeners' dataset. The curves are depicted regarding to one and two second window as well as the combinations of audio/video features

validation results, it may imply that video information is more powerful, but it is more person-dependent. On the other hand, audio information contains more general characteristics among different subjects.

5 Conclusions and Future Work

This paper reports the exploration in using RNN models for the generation of micro listening behaviors for virtual agents. The models are trained on an active listening data corpus which features elderly speakers talking with young active listeners and was collected by ourselves. We compared the performance of MLP, GRU, and LSTM networks and the results show that the reactions of listeners can be trained to some degree merely from the speaker's behaviors. Also, GRU is confirmed to be most effective among the three.

The numeric-wise performance of tested models was not bad, however, it is hard to say that this performance is good or not. From the aspect in learning such person-dependent human behaviors, we would say the accuracy is quite

high. However, from the subjective observation when interacting with such an agent, the perceived performance could be a totally different story. A serious investigation based on subjective perception is required in the future. The proposed model generates facial expressions, head movements, and postures but does not generate gaze which perform an essential role in communication. Unlike facial expressions or head movements like nodding which have absolute meaning, numeric gaze direction values themselves do not reveal any meaning in communication. Gaze directions need to be interpreted by linking with the position of communication interlocutor's eyes or the objects in the environment. A properly designed gaze model is also required in the future. The proposed model transforms a set of input feature values to one single set feature values (i.e. regression). However, human behaviors should not have one single "correct answer" in one situation. The answer should be an acceptable range rather than a single value. We would also like to explore other techniques like generative adversarial networks (GAN) to derive this "range". Finally, an agent who does not speak cannot really perform active listening tasks, the non-verbal behavior generation model while the agent is speaking as well as the model to determine its utterances are also required in the future.

References

1. Baltrusaitis, T., Ahuja, C., Morency, L.: Multimodal machine learning: a survey and taxonomy. CoRR abs/1705.09406 (2017). http://arxiv.org/abs/1705.09406
2. Chen, M., Wang, S., Liang, P.P., Baltrusaitis, T., Zadeh, A., Morency, L.P.: Multimodal sentiment analysis with word-level fusion and reinforcement learning. In: 19th ACM International Conference on Multimodal Interaction (ICMI 2017), Glasgow, UK, November 2017
3. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. CoRR abs/1406.1078, September 2014. http://arxiv.org/abs/1406.1078
4. Ekman, P., Friesen, W.V., Hager, J.C.: Facial Action Coding System (FACS). Website (2002). http://www.face-and-emotion.com/dataface/facs/description.jsp
5. Hasegawa, D., Kaneko, N., Shirakawa, S., Sakuta, H., Sumi, K.: Evaluation of speech-to-gesture generation using bi-directional LSTM network. In: Proceedings of the 18th International Conference on Intelligent Virtual Agents (IVA 2018), Sydney, Australia, pp. 79–86, November 2018
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
7. Huang, H.H., Fukuda, M., van der Struijk, S., Nishida, T.: Integration of DNN generated spontaneous reactions with a generic multimodal framework for embodied conversational agents. In: 6th International Conference on Human-Agent Interaction (HAI 2018), Southampton, UK, December 2018
8. Huang, H.H., et al.: Toward a memory assistant companion for the individuals with mild memory impairment. In: 11th IEEE International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC 2012), Kyoto, pp. 295–299, August 2012

9. Huang, L., Morency, L.-P., Gratch, J.: Virtual rapport 2.0. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 68–79. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23974-8_8
10. Huang, Y., Khan, S.M.: DyadGAN: generating facial expressions in dyadic interactions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, USA, pp. 11–18, July 2017
11. Lausberg, H., Sloetjes, H.: Coding gestural behavior with the NEUROGES-ELAN system. Behav. Res. Methods **41**(3), 841–849 (2009)
12. Otsuka, K., Kasuga, K., Kohler, M.: Estimating visual focus of attention in multiparty meetings using deep convolutional neural networks. In: 20th ACM International Conference on Multimodal Interaction (ICMI 2018), Boulder, USA, pp. 191–199, October 2018
13. Schuller, B., Steidl, S., Batliner, A.: The INTERSPEECH 2009 emotion challenge. In: 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009), Brighton, United Kingdom, September 2009
14. Tickle-Degnen, L., Rosenthal, R.: The nature of rapport and its nonverbal correlates. Psychol. Inq. **1**(4), 285–293 (1990)
15. Wu, J., Ghosh, S., Chollet, M., Ly, S., Mozgai, S., Scherer, S.: NADiA: neural network driven virtual human conversation agents. In: Proceedings of the 18th International Conference on Intelligent Virtual Agents (IVA 2018), Sydney, Australia, pp. 173–178, November 2018